Print

Lesson 11: Estimates and Repeated Sampling - Study Notes

Slide 1:

Parameters

Collecting data on every member of a population is impossible, too costly, or too time-consuming. Therefore, researchers will collect data from a sample and use that resulting sample data to answer questions with regards to the population parameters from which the sample came. Hypothesis testing is one of the statistical procedures used to make **inferences** about a population under study. Although the specific details of a hypothesis test will change with a given situation, the general procedure will remain the same.

Recall: An inference is the process of coming to a logical conclusion based on factual knowledge or evidence.

A parameter refers to a numerical characteristic that defines a population. It is a fixed, unknown value (e.g. average weight of all 30-year-old women in Canada; percentage of voters (p) in Nova Scotia who think the government is doing a good job to control inflation).

Typical population parameters include the mean, the variance, the standard deviation, the area under the probability distribution, and the area between two values of the variable.

For instance, determining an engine's average emission of pollutants, the standard deviation of a bottled-water filling machine, or the variance of the rainfall for the month of March in Uruguay, are all population parameters. Although it is impossible to determine a parameter in all certainty unless we have the data for the entire population, we can predict the value with a certain degree of confidence. Furthermore, some of the most important problems of statistical inference concern the appraisal of the risks and the consequences to which one might be exposed by making generalisations from sample data. This includes an appraisal of the probabilities of making wrong decisions, the chances of making incorrect predications, and the possibility of obtaining estimates that do not lie within permissible limits.

Slide 2:

Estimation

The most obvious way to determine a population parameter is through its corresponding sample statistic. For instance, if we wanted to determine the mean weight of Concordia University students, we may find the mean from a random sample of 100 students (55 kg). The value from the sample serves as a **point estimate** for the actual, unknown population mean. Chances are the value from the sample, although close to the parameter value, is not equal to it. Therefore, we may employ another technique known as **interval estimation**. This procedure involves the establishment of a range of values that should contain the actual population value. This range is always centered on the point estimate. For instance, the mean weight of Concordia Students has a better chance of being between 52 and 58 kg.

Point estimation is similar, in many respects, to firing a revolver at a target. The estimator, generating point estimates, is analogous to the bullet of the revolver, and the parameter of interest to the bull's-eye. Drawing a sample from the population and estimating the value of the parameter is equivalent to firing a single shot at the target. Sometimes, if conditions are right, you will hit the target dead on. However, it may take several attempts to achieve that result. The stray bullets, although not in the bull's-eye, are the closest estimators we have to approximate the intended target.

Slide 3:

Estimation (Cont'd)

Example 1

Estimating Population Parameters

You want to estimate the mean height of women between the ages of 25 and 29 in a particular city. To do so, you measured the heights of a random sample of 100 women between the ages of 25 and 29, selected at the local gymnasium. The results were:

sample mean = 165 cm
sample standard deviation s = 5 cm

What can you conclude about the heights of all women between the ages of 25 and 29 in this population?

Can the sample statistics (i.e. the sample mean and standard deviation) be used to generate inferences about the population parameters? Why or why not?

Show Answer

Answer:

The sample may not have been representative of the population of women between the ages of 25 and 29 in the city, since it was selected from the local gymnasium. Variables such as socio-economic status, health, and physical condition may have acted as confounding variables in the study.

As the sample is probably not representative of the population, we cannot use these statistics to generate inferences. If the sample had been representative of the population, we would have constructed a confidence interval to estimate the population mean using the given sample statistics. We will learn about confidence intervals a little later in this lesson.

Recall: A confounding variable is a variable that was not accounted for that could jeopardize the reliability of the study. In this example, it is possible that women who go to the gym may be taller than average women.

Slide 4:

Bias

Different samples give different values for sample statistics. By taking many different samples and

calculating a sample statistic for each sample (e.g. the sample mean), you could then draw a histogram of all the sample means. A statistic from a sample or randomized experiment can be regarded as a random variable and the histogram is an approximation of its probability distribution. The term sampling distribution is used to describe this distribution, i.e. how the statistic (regarded as a random variable) varies if random samples are repeatedly taken from the population.

If the sampling distribution is known then the ability of the sample statistic to estimate the corresponding population parameter can be determined.

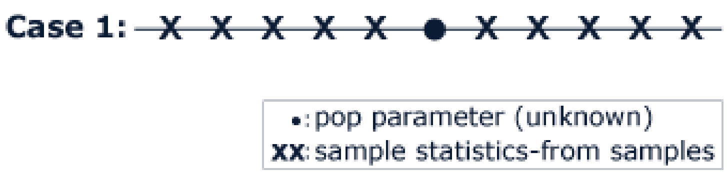
In particular, the sampling distribution determines the expected value and variance of the sampling statistic. If the expected value of the statistic is equal to the population parameter, the estimator is unbiased. If the variance of the statistic is 'small' and it is also unbiased, then an observed statistic is likely to be close to the population parameter.

Bias = distance between parameter and expected value of sample statistics

Slide 5:

Bias (Cont'd)

Subsequently, sample statistics can be classified as shown in the following diagrams.



In case 1, the estimates have a low bias because their mean is near the population parameter, but have high variability due to the fact that they are widely spread out, and a single sample value could be far from the parameter.

These estimates have bias because the expected value (average of sample statistics) is not equal to the parameter. They also have high variability because they are widely spread out.

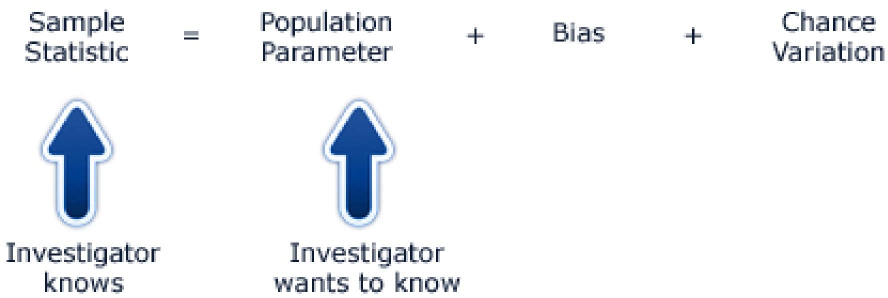
In case 3, the estimates are biased because all of them are systematically higher than the population parameter. The sample statistics have, however, low variability because they are all close together.

In this case (4) the estimates have both low bias and low variability since the mean of the sample statistics is equivalent to the population parameter. Experimental design aims to simultaneously reduce bias and variability by producing a sampling distribution as shown in 4.

Slide 6:

Bias (Cont'd)

Inferences about the characteristics of a population are based on data from a sample of that same population.



If the sample is not representative of the population being studied, the sample statistic may be biased so you cannot use it to make valid inferences about the population parameter.

To minimise bias, the sample should be chosen by random sampling from a list of all individuals in the relevant population. This list is called the sampling frame. It is essential that the random sample be chosen in such a way that each individual in the sampling frame has an equal chance of being selected. This may involve using computer generated random numbers to select the sample.

Slide 7:

Bias (Con't)

Example

Health survey conducted in Lanark County, Ontario.

Study population - all residents of Lanark County, Ontario - aged 25-69 years (approximately 60,000 people).

Sampling frame - electoral roll (some bias has been introduced here because younger people (< 35 years) and migrants are less likely to be on the roll).

Sample selection - sample chosen using computer generated random numbers so each person on the electoral roll in this age group has a 1 in 100 chance of selection. This means that 60,000/100 = 6,000 people were selected for the sample.

Actual sample - includes those who responded to the request to participate in the study.

Non-respondents may differ from the respondents in many ways (e.g. being less healthy) and this could lead to bias in estimates of the proportion of smokers, average weight, etc.

Statistics obtained from a sample that is not representative of the population do not allow us to make accurate inferences about the population parameters. However, even when the sample is representative of the population, the statistics obtained rarely correspond exactly to the population parameters. Nonetheless, they provide us with the best possible estimation.

Slide 8:

The Central Limit Theorem and the Sampling Distribution of the

Sample Mean

The sampling distribution of the sample mean $\bar{x}$ has unique properties. If a random sample of n observations is drawn from a population with a mean μ and a standard deviation σ, the **sampling distribution** of $\bar{x}$ (the mean of a bunch of sample means) will have a mean of μ (the same as the mean of the sample's population) and a standard deviation equal to the population standard deviation divided by the sample size (σ / sqrt n). The most important property, however, is a result known in statistics as the Central Limit Theorem. This theorem states that when the sample size n is large, the sampling distribution of the sample mean will be approximately normally distributed. A "large" sample size is universally considered as an n of 30 or more measurements. The approximation will become increasingly accurate (closer to the actual, unknown value) as n becomes larger and larger.

The Central Limit Theorem

The sampling distribution of sample means will more closely resemble the normal as sample size increases.

Slide 9:

The Central Limit Theorem and the Sampling Distribution of the

Sample Mean

Example 1

Consider a population (e.g., a game) that consists of only 4 possible scores: 2, 4, 6, and 8. We make use of this population to identify all the possible outcomes of n = 2 and construct a distribution of those results.

We would like to determine if:

- The mean of sample = the mean of population.
- The standard deviation of sample approximately = the standard deviation of the population divided by the square root of the sample size.
- The sampled population has a normal distribution.

style="TEXT-ALIGN: center">Sample	1st scores	2nd scores	Sample Mean $\bar{x}$
1	2	2	2

2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

Possible samples of n = 2.

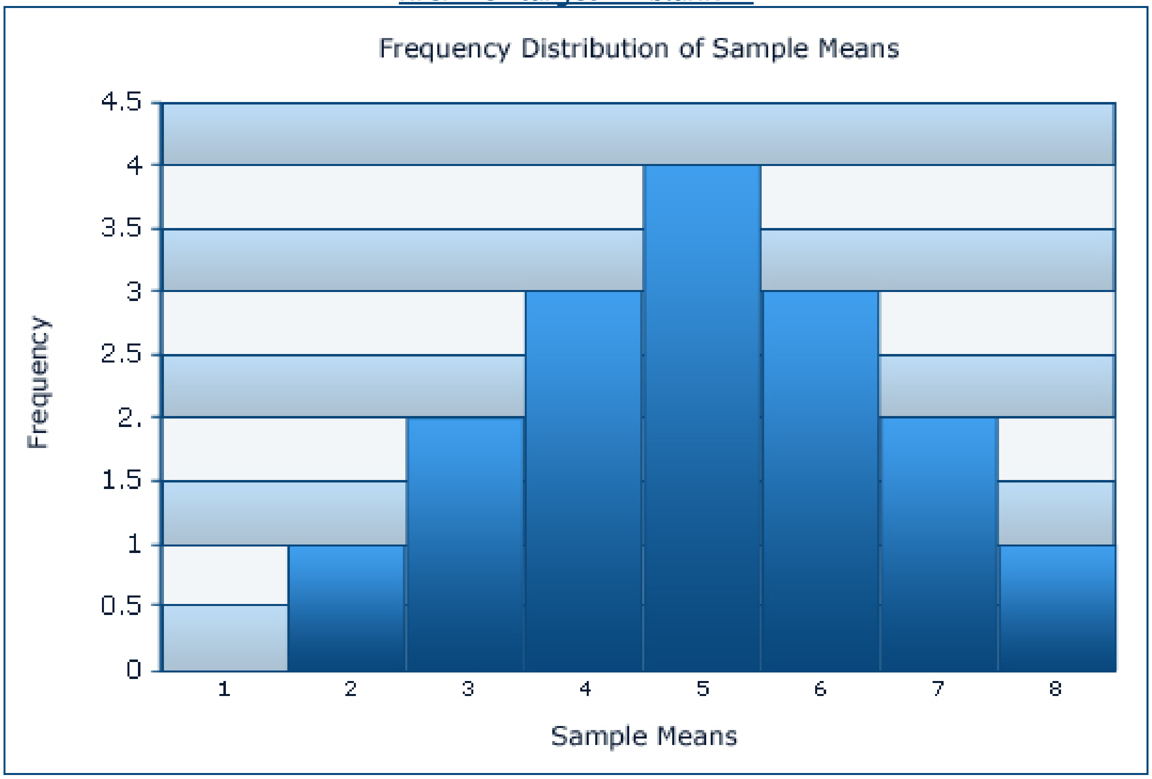
The table above can be simplified as follows:

Sample Mean $\bar{x}$	**style="TEXT-ALIGN: center">Frequency**
1	0
2	1
3	2
4	3
5	4
6	3
7	2
8	1

Slide 10:

Frequency Distribution of Sample Means

href="s" target="_blank">



The distribution of sample means for n = 2.

Answer:

Mean of the sample = mean of the population = **5**

SD of sample = SD of population = **1.63**

The distribution is normal (see figure above).

Slide 11:

Recap

In order to determine the parameter values of a population, it is much more convenient, and sometimes necessary, to make estimates based on a random sampling of the population under study. However, in order to get a proper estimation, several factors must be considered.

- An ideal estimate would stem from the statistics of a representative sample.
- The presence of bias in a sample will distort the predictions made about the population under study.
- The sampling distribution of sample means will gradually resemble the normal

distribution as the sample size increases.

You can post a message online in your discussion folder any time you have something to share with your discussion group concerning the current lesson. Simply click Discussion Board or use the menu at the top of the screen.

Next lesson: Confidence Intervals and Coefficient